Gradient boosting machine assisted approximate Bayesian computation for uncertainty analysis of rainfall-runoff model parameters

F.R. Bennett

Queensland Department of Natural Resources, Mines and Energy, Dutton Park
Email: frederick.bennett@des.qld.gov.au

Abstract: Bayesian inference is a well-regarded approach for diagnostic model evaluation that is often applied to hydrological models to constrain parameters and estimate uncertainty within a statistical framework. Typically, Bayesian uncertainty analysis is carried out using the Generalised Likelihood Uncertainty Estimation (GLUE) or Markov Chain Monte Carlo (MCMC) sampling. Approximate Bayesian Computation (ABC) is alternative set of likelihood-free Bayesian methods that have been gathering interest in many fields including astrophysics, population genetics and biology.

The main appeal of ABC is that is the requirement for a formal likelihood function is replaced by one or more summary statistics that compare the simulated model to the observed data. ABC works in situations where an analytical likelihood function is either unavailable or intractable. Instead of evaluating the likelihood function, ABC only has to be able to sample from the likelihood function in an empirical fashion. This broadens the class of problems to which statistical inference can be applied.

In practice, the appeal of the ABC method is limited somewhat due to its requirement for a large number of model evaluations. Because ABC is essentially a rejection sampling method, when the overlap between the prior and the posterior is poor, the sampling efficiency can be very low and it may be necessary for hundreds of thousands or even millions of model evaluations to be run to collect an appropriate number of accepted samples to construct a statistically informative posterior. If the model runtime is significant, ABC rejection sampling can easily be rendered impractical. In this paper a hybrid method is developed that serves to retain the flexibility of ABC while drastically reducing the computational effort required. The first component of the hybrid approach is to employ Sequential Monte Carlo sampling (SMC) to improve the sampling efficiency and reduce the total number of samples required. Secondly, the primitive model is replaced by a surrogate model that can accurately reproduce the results of the original model at a fraction of the computational cost. In this case, XGBoost, a gradient boosted regression tree machine learning algorithm, is used to construct the surrogate models.

Employed together, SMC-ABC and XGBoost trained surrogate models offer an accurate and efficient framework for model parameter inference and uncertainty analysis. As a demonstration, the proposed method is applied to a four parameter GR4J distributed rainfall runoff model to estimate marginal model parameter probability density functions for parameter identification and uncertainty analysis.

Keywords: *Uncertainty analysis, conceptual model, Bayesian inference*

1. INTRODUCTION

It serves no useful purpose to imagine that environmental models are mechanistically faithful representations of the processes that they seek to emulate. In the case of conceptual rainfall-runoff models (CRR), structural deficiencies along with noisy observational forcing and response data are significant contributors to uncertainty of model predictions. Given this practically unavoidable condition, it is important to be able to quantify the magnitude of uncertainty of CRR models to properly assign a justifiable level of confidence in their results.

The most common approach to quantifying CRR model uncertainty is Bayesian inference as formalised in Bayes' Theorem which can be expressed as

$$p(\theta|D) \propto p\big(D\big|M(\theta)\big)p(\theta) \tag{1}$$

where D is the available data, M is the model and θ is a set of model parameters. Bayes' Theorem states that the desired posterior parameter probability given the available data, $p(\theta|D)$, is proportional to the product of a prior parameter probability, $p(\theta)$, and the likelihood for the data $p\big(D\big|M(\theta)\big)$ given the model evaluated with parameters θ. The elegance of Bayes' theorem belies the fact that although the prior can usually be easily decided upon based on some understanding of the behaviour of the model, the likelihood is typically unknown, at least in an analytical form. In practice, assumptions about the form of the likelihood function are often made and formal statistical models are employed as a convenience. It can be argued, however, that the reliance on these assumptions can be problematic, particularly when model residual errors are correlated, nonstationary and non-Gaussian.

Approximate Bayesian Computation (ABC) is a so-called likelihood-free method for Bayesian inference which eliminates the requirement for a formal likelihood function. When provided with a sufficient summary statistic describing qualities of the model, ABC empirically estimates the Bayes' posterior (Beaumont 2019). The ABC algorithm imposes a performance-threshold based on the summary statistic that serves to identify and reject regions of the prior parameter space that lead to poorly performing models. If the summary statistic is sufficient and the rejection threshold is suitably assigned, the distribution of accepted parameter combinations sampled from the prior will provide a good estimate of the true Bayesian posterior .

ABC does come with some caveats that may inhibit its practical application. Primarily, the computational effort required for rejection sampling may easily become unmanageable. The classical ABC rejection sampler can be highly inefficient and the number of model runs required to build an approximation of the posterior density function can number in the hundreds of thousands. This paper investigates two strategies for reducing the effective computation burden of ABC sampling by orders of magnitude. The first is to improve the sampling algorithm to reduce the total number of model evaluations required. This is achieved through adaptive sequential Monte Carlo (SMC) sampling (Beaumont et al. 2009). The SMC-ABC algorithm iteratively adapts the summary-statistic rejection threshold while reducing the rejection rate by increasing sampling from the high density region of the posterior (Price et al. 2018).

Although SMC-ABC delivers welcome efficiency gains over simple rejection sampling, more drastic computation efficiencies can be realised by replacing the computationally demanding primitive model with an emulator or surrogate model. Surrogate models are evaluated rapidly and can be trained to reproduce the required summary statistics as a function of model parameters with satisfactory accuracy. In the present study, extreme gradient boosted regression trees (XGBoost) (Chen and Guestrin 2016) are trained for this purpose.

Employed together, SMC-ABC and XGBoost trained surrogate models offer an accurate and efficient framework for model parameter inference and uncertainty analysis. As a demonstration, the proposed method is applied to a four parameter GR4J distributed rainfall runoff model to estimate marginal model parameter probability density functions for parameter identification and uncertainty analysis.

2. METHODS AND RESULTS

2.1. ABC Algorithm

The aim of ABC is to sample from a posterior that is conditioned by observational data without any reference to likelihood functions. This is achieved by replacing the likelihood term in (1) by a comparison between observed (D) and simulated data expressed in terms of a summary statistic as shown below in (2)

$$p(\theta|D)_\epsilon \propto \int M(\boldsymbol{y}|\theta)\mathbb{1}\big(\rho\big(S(\boldsymbol{y}),S(D)\big)\leq\epsilon\big)p(\theta)d\boldsymbol{y} \quad (2)$$

where M is a model that given parameters θ generates the simulated data $\boldsymbol{y}$. $\rho\big(S(\boldsymbol{y}),S(D)\big)$ is a distance metric comparing the summary statistic evaluated for the simulated and observed data and ϵ is the acceptance tolerance. The distance metric may take in various forms but a simple Euclidean norm $\|S(\boldsymbol{y})-S(D)\|$ is quite often a suitable choice. The basic principle of ABC is that if the summary statistic is sufficient and ϵ is small then sampling from $p(\theta|D)_\epsilon$ should give a good approximation of the true posterior $p(\theta|D)$ (Marjoram et al. 2003; Sisson et al. 2007).

One of the first applications of ABC was in the field of population genetics where the procedure depicted in the pseudocode of Algorithm1 was introduced (Pritchard et al. 1999). Although this routine is a perfectly adequate for sampling from $p(\theta|D)_\epsilon$ if ϵ is small and a high degree of mismatch between the prior and posterior distributions exists, the acceptance rate can be very low. When the runtime time for the model is significant, inefficient sampling may effectively render Algorithm 1 impractical.

ABC with Sequential Monte Carlo sampling (ABC-SMC) (Toni et al. 2009) is a type of importance sampling (Del Moral et al. 2006; Liu and Chen 1998) that improves on the rudimentary ABC algorithm by enhancing sampling efficiencies. The resulting ABC-SMC strategy is to sample iteratively from a sequence of distributions $p(\theta|D)_{\epsilon,j}$ for decreasing ϵ_j. The ABC-SMC procedure is outlined in Algorithm 2 shows that as the acceptance tolerance is gradually decreased with each iteration, the algorithm learns from previous iterations where in parameter space the higher density region of the posterior distribution exists. For each iteration, particle importance weights are calculated using an adaptive transition kernel, q, that depends on the variance of the particle population from the previous iteration. The weighting scheme has the function of minimising the Kullback-Leibler distance between the intermediate proposal and posterior density functions. Minimising the Kullback-Leibler distance will maximise the acceptance rate (Beaumont et al. 2009; Turner and Van Zandt 2012). An average 50-fold improvement in sampling efficiency by using ABC-SMC compared to ABC-Rejection has been reported (Toni et al. 2009).

Algorithm 1 ABC-Rejection

Set tolerance ϵ
for i = 1,…, N **do**
  while $\rho\big(S(\boldsymbol{y}_i)-S(D)\big)>\epsilon$ **do**
    Sample θ_i^* from the prior $p(\theta)$
    Simulate a dataset $\boldsymbol{y}_i{\sim}M(\theta_i^*)$ and evaluate summary statistic $S(\boldsymbol{y}_i)$
  end while
  set $\theta_i=\theta_i^*$
end for

Algorithm 2 ABC-SMC

At iteration j = 0,
Set $\epsilon_0 \leftarrow \epsilon$
for i = 1,…,N **do**
  while $\rho\big(S(\boldsymbol{y}_i)-S(D)\big)>\epsilon_0$ **do**
    Sample θ_i^* from the prior $p(\theta)$
    Simulate a dataset $\boldsymbol{y}_i{\sim}M(\theta_i^*)$ and evaluate summary statistic $S(\boldsymbol{y}_i)$
  end while
  set $\theta_{i,0}=\theta_i^*$
  set weights for each particle $w_{i,0}\leftarrow 1/N$
end for
set the covariance $\sigma_0^2\leftarrow 2\mathrm{Cov}\big(\theta_{1:N;0}\big)$
At iteration j > 0
for j = 1,…, J **do**
  Set to $\epsilon_j \leftarrow P_{70}\,\rho\big(S(\boldsymbol{y}_i)-S(D)\big)$
  for i = 1,…,N **do**
    while $\rho\big(S(\boldsymbol{y}_i)-S(D)\big)>\epsilon_j$ **do**
      Sample random θ_i^* from previous iteration $\theta_i^*{\sim}\theta_{1:N;t-1}$
      Perturb θ_i^* by sampling $\theta_i^{**}{\sim}N\big(\theta_i^*,\sigma_{j-1}^2\big)$
      Simulate a dataset $\boldsymbol{y}_i{\sim}M(\theta_i^{**})$ and evaluate summary statistic $S(\boldsymbol{y}_i)$
    end while
    set $\theta_{i,j}\leftarrow\theta_i^{**}$
    set $w_{i,j}\leftarrow\frac{\pi(\theta_{i,j})}{\sum_{k=1}^{N}w_{k,j-1}q\big(\theta_{k,j-1}|\theta_{i,j},\sigma_{j-1}^2\big)}$
  end for
  Set the covariance $\sigma_j^2\leftarrow 2\mathrm{Cov}\big(\theta_{1:N;0}\big)$
end for

2.2. XGBoost

Even with the significant sampling efficiency gains that can be realized through sequential Monte Carlo sampling, the number of model evaluations required to satisfactorily populate a posterior probability distribution through ABC can still be prohibitively high. In the experience of the GBR catchment modelling team, a single catchment model run can take anywhere between a few minutes to two hours to run. Replacing the primitive catchment model with a well-trained emulator or surrogate model has the potential to reduce run times by several orders of magnitude. In the present case, the focus is on identifying a suitable emulator for estimating ABC summary statistics as a function of model parameters. Machine learning methods have been widely applied as emulators of deterministic models and include neural networks, Gaussian processes,

polynomial regression, polynomial chaos expansions, radial basis functions and multivariate adaptive regression splines, random forests and generalised additive models to name some popular options (see Asher *et al*., 2015 for a review of surrogate models applied to groundwater modelling).

Gradient boosted decision tree (GBDT) algorithms are a family of machine learning techniques that have emerged as a powerful tool for classification and regression applications (Roy et al. 2019). First introduced by Friedman (Friedman 2001), gradient boosting is a method for synthesising a statistical model from an ensemble of weak models. Put simply, gradient boosting progressively fits a sequence of models to gradually improve the estimate of the dependent variable. This is distinctive from the popular bagging technique where an ensemble of predictors are determined independently of each other. The basic idea behind the gradient algorithm is to create a series of new base-learners that are correlated with the negative gradient of the loss function that is associated with the entire ensemble. For GBDTs, base-learners are represented by decision trees.

Several modern open-source GBDT toolkits have become available to popular modelling platforms in recent years. CatBoost (Prokhorenkova et al. 2018) and LightGBM (Ke et al. 2017) both have considerable support in the machine learning community, but for the current study the XGBoost package (Chen and Guestrin 2016) has been chosen.

2.3. XGBoost Assisted ABC Method

ABC demands an appropriate summary statistic to be defined to evaluate not only the goodness-of-fit of the model to the observed data, but also provide some information about the system behaviour (Vrugt and Sadegh 2013). The proposed composite statistic in (3) includes the correlation coefficient, *r*, which directly compares paired values in the observed and simulated time series as well as the first three statistical moments that give a comparison of the centre, width and shape of the distribution of values across the entire time series. μ, σ and γ are respectively the mean, standard deviation and skewness of the time series data distribution and the subscripts *m* and *o* are for the modelled and observed data. *a*, *b*, *c* and *d* are weighting coefficients that can be adjusted to emphasise or attenuate the influence of each component of the summary statistic.

$$S = \sqrt{a(r-1)^2 + b\left(\frac{\mu_m}{\mu_o} - 1\right)^2 + c\left(\frac{\sigma_m}{\sigma_o} - 1\right)^2 + d\left(\frac{\gamma_m}{\gamma_o} - 1\right)^2} \quad (3)$$

The form of (3) results in $S(D) = 0$ so $\rho\big(S(\boldsymbol{y}_i) - S(D)\big)$ conveniently becomes $S(\boldsymbol{y}_i)$ in cases where the Euclidean distance metric is used.

Although it would be possible to build a single surrogate model to estimate the composite statistic in (3), in practice more accuracy and flexibility has been achieved by building individual surrogate models for each of the components of the summary statistic r, μ_m, σ_m and γ_m and then using these estimates to calculate the composite summary statistic S.

2.4. Blencoe Creek Catchment Case Study

Blencoe Creek is a small tributary to the Herbert River in North Queensland. A catchment area of 223.7km^2 reports to the Blencoe Falls hydrologic reference station GS116010A (Zhang et al. 2014). The catchment is in pristine condition with no modifications or regulated water use. A four parameter GR4J lumped hydrological model (Perrin et al. 2003) was built using the Source modelling platform using daily rainfall and evapotranspiration forcing data sourced from SILO (Jeffrey et al. 2001).

Table 1. GR4J model parameters and ranges

Name	Unit	Range	Description
X1	mm	1.0 – 1500.0	Maximum capacity of production store
X2	mm	-10.0 – 5.0	Groundwater exchange coefficient
X3	mm	1.0 - 500	1-day-ahead maximum capacity of the routing store
X4	day	0.5 – 4.0	Time base of the unit hydrograph, UH1

The GBDT summary statistic emulator was constructed from training and evaluation data chosen using a Sobol sequence sampling design of experiment (Sobol 1967). 4096 GR4J models were evaluated distributed over the parameter ranges given in Table 1. As summarised in section 2.3, based on the training set of data, surrogate

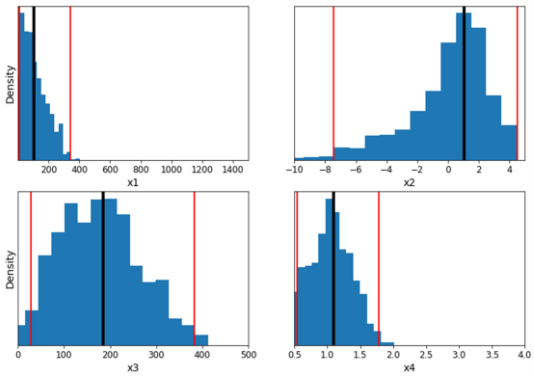

Figure 1. Posterior marginal parameter density plots for the GR4J model. The thick black line represents the expectation value of the distributions and the red lines are placed at the 95% confidence limits.

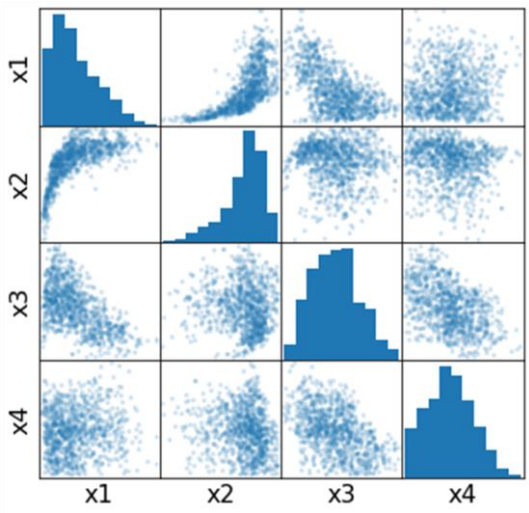

Figure 2. Parameter correlation plots

models for each summary statistic component, r, μ_m, σ_m, and γ_m were constructed. The XGBoost python package (Anon 2019) was used to implement the GBDT machine learning algorithm. XGBoost hyperparameters were optimised using a Bayesian search method based on n-fold cross validation scores (Snoek et al. 2012). Overfitting of the surrogate models was avoided by using early stopping (Zhang and Yu 2005) to limit the total number iterations in the GBRT gradient descent.

The marginal posterior distributions for Blencoe Creek GR4J model parameters are shown in Figure 1. Each parameter is well identified with well resolved unimodal distributions. Ideally, parameters would not only be identifiable but also independent whereby each parameter value would not influence that of another parameter (*i.e.* not exhibit cross-correlation). Bivariate parameter correlation plots in Figure 2 reveal a degree of positive correlation between X1 and X2 where a large portion of the X1-X3 plots is empty and thus indicate regions of non-behavioural parameter combination. There are weaker negative correlations between parameters X1 and X3 as well as X4 and X3 suggesting that these pairs of samples may tend to counterbalance each other in the streamflow model results.

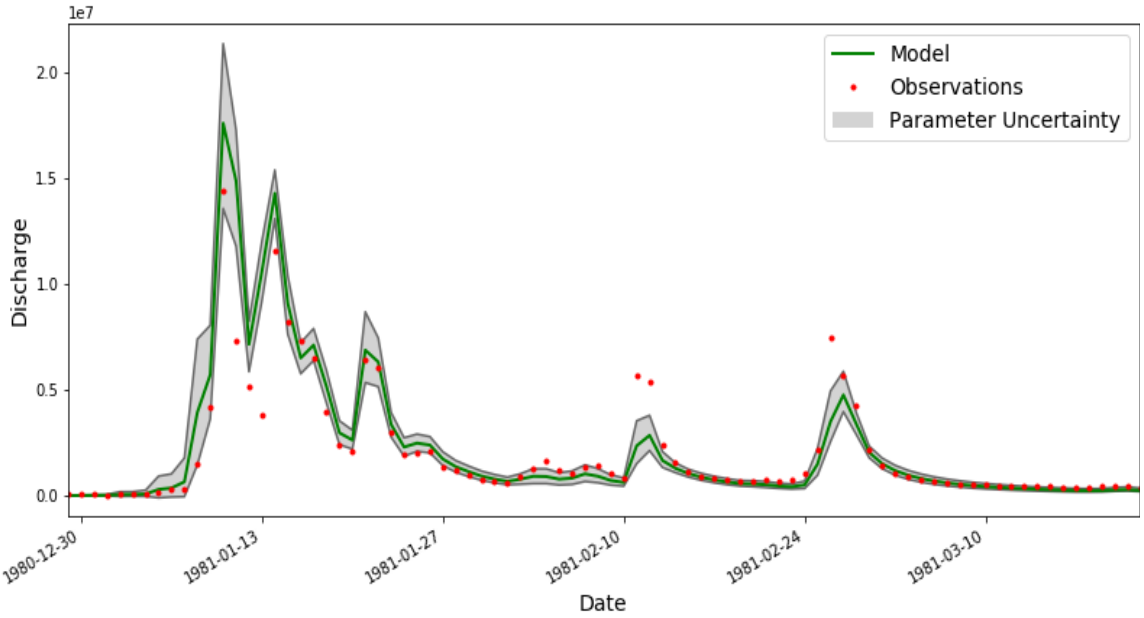

Figure 3. Predictive uncertainty plot for a selected time series range including the 1981 wet season.

The expectation values of each parameter (taken as the mean of the marginal posterior distribution) along with the 95% confidence limits were used to synthesise the flow time series shown in Figure 3. A strong relationship

between the modelled and observed streamflow data is evident with most of the observational data sitting within or on the perimeter of the predictive uncertainty interval.

3. DISCUSSION AND CONCLUSIONS

The application of a combination of machine learning regression methods and SMC-ABC to the parameter identification and uncertainty analysis of a rainfall runoff model has been demonstrated. This combined approach is designed to overcome practical sampling inefficiency problems associated with the ABC-Rejection algorithm. This is particularly pertinent in the case of GBR water quality models where model runtimes can range between 20 minutes and 2 hours per simulation on a typical workstation. Depending on the divergence between the parameter prior and conditional posterior distributions ABC may require hundreds of thousands or even millions of simulation runs in order to accumulate a suitably sized set of accepted samples for statistical analyses of the posterior. In the case presented in a companion paper where water quality parameters have been studied (Baheerathan and Bennett 2019) it is estimated that the computational effort for parameter inference and uncertainty analysis has been reduced by four orders of magnitude over ABC-Rejection sampling using the primitive model.

In principal, once a suitable surrogate model has been built, it could be applied to a variety of Monte Carlo methods but generally speaking, some of the interesting features of the SMC-ABC approach to model parameter inference include;

- Along with rejection sampling, the SMC-ABC sampler is not prone to becoming stuck in low probability regions of the posterior distribution surface.
- Unlike rejection sampling (including GLUE), SMC-ABC is not burdened by sampling inefficiencies caused by large divergences between the prior and posterior distributions.
- The empirical nature of the SMC-ABC sampler easily accommodates complex posterior distribution functions.
- In contrast to MCMC samplers, SMC-ABC samples are not correlated and a burn in phase is not required by the algorithm.

When a suitable surrogate model for estimating the summary statistic is available, SMC-ABC becomes an efficient and highly flexible method for Bayesian parameter inference and uncertainty analysis.

REFERENCES

Anon (2019) *XGBoost Python Package*. https://xgboost.readthedocs.io/en/latest/python/index.html Accessed 23 July 2019.

Asher, M.J., Croke, B.F.W., Jakeman, A.J. and Peeters, L.J.M. (2015) A review of surrogate models and their application to groundwater modeling. *Water Resources Research* 51 (8), 5957–5973.

Baheerathan, R. and Bennett, F.R. (2019) Bayesian uncertainty analyses of Great Barrier Reef catchment water quality model parameters without likelihood assumptions. *MODSIM2019* Canberra, Australia.

Beaumont, M.A. (2019) Approximate Bayesian Computation. *Annual Review of Statistics and Its Application* 6 (1), 379–403.

Beaumont, M.A., Cornuet, J.-M., Marin, J.-M. and Robert, C.P. (2009) Adaptive approximate Bayesian computation. *Biometrika* 96 (4), 983–990.

Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* San Francisco, California, USA: ACM Press785–794.

Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (3), 411–436.

Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29 (5), 1189–1232.

Jeffrey, S.J., Carter, J.O., Moodie, K.B. and Beswick, A.R. (2001) Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software* 16 (4), 309–330.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. in I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan & R Garnett (editors), (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30* Curran Associates, Inc. 3146–3154.

Liu, J.S. and Chen, R. (1998) Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association* 93 (443), 1032–1044.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100 (26), 15324–15328.

Perrin, C., Michel, C. and Andréassian, V. (2003) Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology* 279 (1), 275–289.

Price, L.F., Drovandi, C.C., Lee, A. and Nott, D.J. (2018) Bayesian Synthetic Likelihood. *Journal of Computational and Graphical Statistics* 27 (1), 1–11.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. and Feldman, M.W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16 (12), 1791–1798.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. in S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi & R Garnett (editors), (2018) CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31* Curran Associates, Inc. 6638–6648.

Roy, A., Qureshi, S., Pande, K., Nair, D., Gairola, K., Jain, P., Singh, S., Sharma, K., Jagadale, A., Lin, Y.-Y., Sharma, S., Gotety, R., Zhang, Y., Tang, J., Mehta, T., Sindhanuru, H., Okafor, N., Das, S., Gopal, C.N., Rudraraju, S.B. and Kakarlapudi, A.V. (2019) Performance Comparison of Machine Learning Platforms. *INFORMS Journal on Computing* 31 (2), 207–225.

Sisson, S.A., Fan, Y. and Tanaka, M.M. (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 104 (6), 1760–1765.

Snoek, J., Larochelle, H. and Adams, R.P. (2012) Practical Bayesian Optimization of Machine Learning Algorithms. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2* USA: Curran Associates Inc.2951–2959.

Sobol', I.. (1967) On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* 7 (4), 86–112.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M.P.H. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface* 6 (31), 187–202.

Turner, B.M. and Van Zandt, T. (2012) A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology* 56 (2), 69–85.

Vrugt, J.A. and Sadegh, M. (2013) Toward diagnostic model calibration and evaluation: Approximate Bayesian computation: Approximate Bayesian Computation. *Water Resources Research* 49 (7), 4335–4345.

Zhang, S.X., Bari, M., Amirthanathan, G., Kent, D., MacDonald, A. and Shin, D. (2014) Hydrologic reference stations to monitor climate-driven streamflow variability and trends. *Hydrology and Water Resources Symposium 2014* 1048.

Zhang, T. and Yu, B. (2005) Boosting with early stopping: Convergence and consistency. *The Annals of Statistics* 33 (4), 1538–1579.